Atom depth in protein structure and function

Alessandro Pintar[1], Oliviero Carugo[1,2] and Sándor Pongor[1]

[1]Protein Structure and Bioinformatics Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), AREA Science Park, Padriciano 99, 34012 Trieste, Italy
[2]Department of General Chemistry, University of Pavia, Viale Taramelli 12, 27100 Pavia, Italy

Atom depth, originally defined as the distance between a protein atom and the nearest water molecule surrounding a protein, is a simple but valuable geometrical descriptor of the protein interior. It can be easily computed from the 3D structure of a protein, thus complementing the information provided by the calculation of the solvent accessible surface area and buried surface area. Depth has been found to be correlated with several molecular, residue and atomic properties, such as average protein domain size, protein stability, free energy of formation of protein complexes, amino acid type hydrophobicity, residue conservation and hydrogen/deuterium amide proton exchange rates.

Although the functional properties of a protein, including the enzymatic activity, the binding of small ligands, the interaction with other proteins and the whole battery of post-translational modifications that play a key role in recognition and signaling are all related to its surface properties, its ability to fold correctly into the native state and to maintain this fold stably, which ultimately assures its function, is related to the features of the protein interior. It has long been recognized that the driving force for a polypeptide chain to fold into a compact globular state in an aqueous environment is mainly produced by the exclusion of water molecules and the formation of a hydrophobic core, whereas the thermodynamic stability of the native structure strongly depends on the packing interactions occurring between residues forming this core. A precise and accurate description of the protein core is thus necessary not only for the understanding of protein folding and stability but also for protein engineering and *de novo* protein design. A geometric parameter, 'depth', was devised to gain insight into the protein interior. We describe here how depth can be defined and computed, and how it has been applied in the analysis of protein structure and function.

Depth calculation methods

Depth can be defined as the distance between a protein atom and the nearest water molecule surrounding the protein [1] (Box 1), and methods of different complexity have been proposed to place the water molecules around the protein and to calculate this parameter [1,2]. In the

'nearest hypothetical water molecule' method, the protein is placed in a 3D lattice containing water molecules [1]. In a slightly different approach [2], the protein molecule is placed in a water box obtained from a Monte Carlo simulation, the dynamics of the protein are approximated through sequential rotations and translations, and the average distance of every atom from the nearest water molecule computed. In this approach, the water molecules that are found both in cavities and in clefts or surface grooves are explicitly removed. In what has been described as the most accurate, but also the most computationally intensive method, the average distance between a protein atom and a water molecule is calculated during a molecular dynamics simulation of the solvated protein [2]. Atom depth can also be defined as the distance of an atom from the molecular surface or the solvent accessible surface [2]. A dot surface must then be calculated [3], as well as the distance of every atom from its closest surface dot. In a more recent approach, depth is defined as the distance of a buried atom from its closest solvent-accessible protein neighbor [4], solvent-accessible atoms being identified through the rolling sphere algorithm [5,6]. It is likely that the interaction of a water molecule with a protein cannot be described accurately by a rolling sphere because this interaction is mainly dictated by hydrogen bonding, and hence by strongly directional forces [2]. Moreover, the 'nearest solvent-accessible atom' approach introduces some arbitrary parameters: the radius of the probe used to identify solvent-accessible atoms (1.4 Å as default) and the cut-off used to define an atom as 'accessible' (0.0 Å² as default). The claimed advantage is represented by the removal of explicit water molecules as well as surface dots from the calculation, with a significant reduction in computational time. It should be noted that in the 'nearest solvent-accessible atom' approach, the simplification of the algorithm leads to some loss of information for surface atoms, for which depth is zero by definition.

Independently of the definition and the algorithm used for the calculation, the structural features of the protein interior can be easily visualized using the depth parameter. In Figure 1, depth was calculated for the structure of interferon-α2a [7], and atoms colored according to their depth values [4]. Despite the apparent symmetry of the molecule, there is a strong unevenness in the distribution of the core atoms, the most deeply buried atoms being all

<div style="border:1px solid #000;padding:1em;background:#fffbe6;">

Box 1. Defining 'depth'

Slightly different definitions of 'depth' have been used and, accordingly, different algorithms have been developed to calculate it. Depth can be defined as the distance between a protein atom and the nearest water molecule surrounding the protein (Figure Ia), as the distance between a protein atom and the molecular (or solvent-accessible) surface (Figure Ib), and as the distance between a protein atom and its closest solvent-accessible neighbor (Figure Ic).
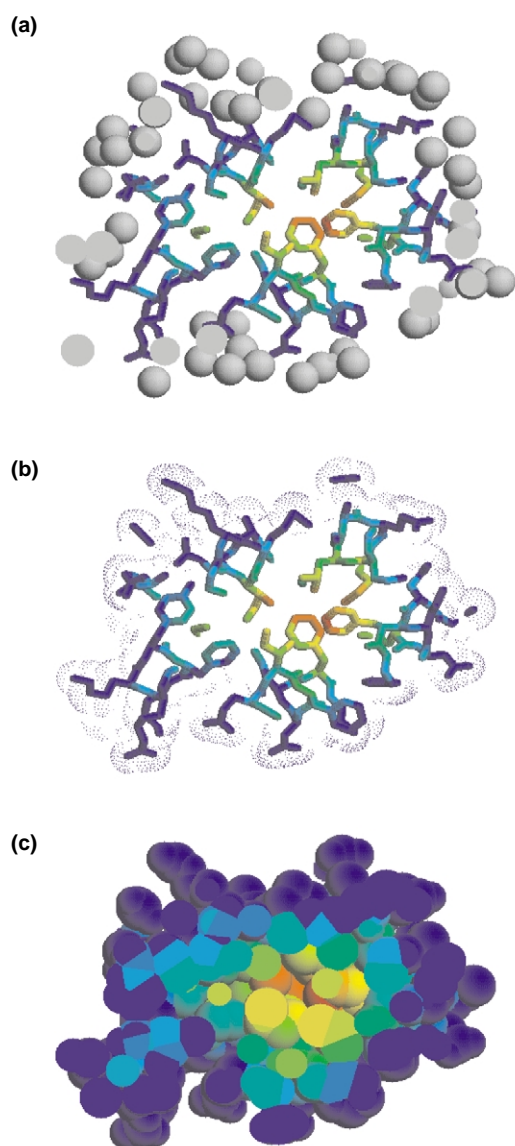
(a)

(b)

(c)

Figure I. Cross section of the structure of interferon-α2a [7], perpendicular to the axis of the helix bundle. In **(a)**, water molecules are shown as gray spheres and the protein as sticks. In **(b)**, the molecular surface is represented by dots. In **(c)**, protein atoms are shown as CPK spheres. In all views, atoms (bonds) are colored according to depth; the color scale goes from red for the deepest buried to blue for the surface atoms.

</div>

localized in a cluster around L18. Because depth is an atomic property, it can be easily handled: backbone, side chain and mean residue values can be computed and plotted against the residue number (Figure 1).

Depth values calculated by the 'nearest solvent accessible atom' method [4] and by the 'nearest water molecule' approach [2] for the same protein structure show a close linear correlation (R = 0.93 for interferon-α2a). However, as one might expect, values calculated using the first method [4] are systematically smaller by ~3.5 Å then those calculated by the 'nearest water molecule' method [2]. This difference arises from the different definitions of 'depth' (see Box 1), and roughly corresponds to twice the van der Waals radius of a non-hydrogen atom.

Evaluating the accuracy of these different approaches is at present a difficult task. It has been proposed that the position of water molecules in high-resolution crystal structures can be used as a benchmark [2]. However, the number of visible crystallographic water molecules is usually very limited [8]; furthermore, it is questionable whether the dynamic behavior of both the protein and the water shell can be effectively represented in a crystal. It can be argued that hydrogen/deuterium (H/D) exchange rates, as measured in solution by nuclear magnetic resonance (NMR) [9], can provide a reliable test; however, there are also serious limitations. Common NMR techniques can directly monitor only protons that exchange with the solvent, most significantly backbone NH amides, for which exchange rates strongly depend on hydrogen bonding, and not only on depth.

One might question why we need yet another parameter to describe the physical properties of a protein. The solvent-accessible area [5] has been the most successfully and widely used for this purpose, and several algorithms have been designed to calculate this parameter from the 3D structure [6], or to predict it from the amino acid sequence only [10]. However, solvent accessibility does not provide useful structural information on atoms and residues that are buried within the protein interior. In a similar way, methods aimed at the calculation of the occluded or buried surface area [11] cannot distinguish residues that are buried but close to the protein surface from those that are deeply buried in the protein core. The same limitations apply to the calculation of the number of contacts made by an atom with its neighbors in a 3D protein structure. As an example, the residue solvent accessibility, the buried surface area and the number of contacts have been calculated for interferon-α2a and plotted against the mean residue depth (Figure 2). Whereas for surface residues all of these parameters show a high sensitivity and a close correlation, sensitivity clearly decreases when we move towards the protein interior [2,12]. Depth is thus a geometric parameter that can complement the information provided by the solvent accessible/buried surface area and by the number of contacts, and has proven to be a valuable descriptor in several applications, which are summarized below.

Depth in protein structure and function analysis

The molecular level

It is well known that the 'size' of globular proteins and protein domains does not increase indefinitely with the number of residue in the chain. An unlimited increase in
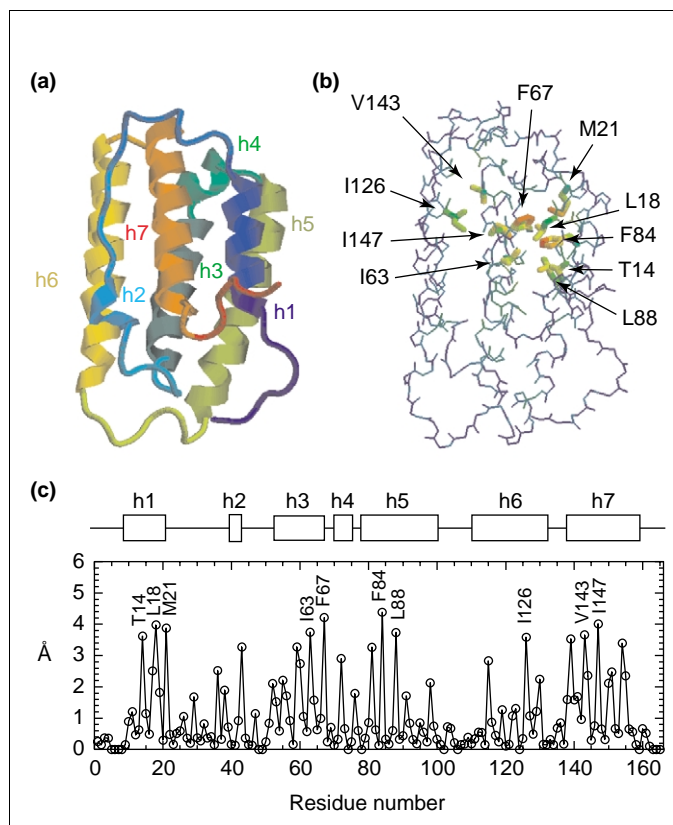
Figure 1. The structure of interferon-α2a [7]. **(a)** Schematic model, colored according to the sequence number from blue (N terminus) to red (C terminus), and in the same orientation as the wireframe model **(b)**, colored according to atom depth. **(c)** A plot of mean residue depth (Å) versus the residue number, with a sketch of the secondary structure showing helices 1–7. Depth was calculated using DPX [4] and the structure displayed using RASMOL [21]. The ten deepest residues are labeled (one letter code).

the dimension might not be useful because it would slow down the folding process without any significant increase in the solvent-exposed surface, which is the part of a protein that is actively involved in the interaction with other biomolecules. From the calculation of the free energy of unfolding, it has been estimated that the optimal size for a protein domain should be ~100 amino acids [13], and a statistical analysis of domain size distribution came to the

same result [14]. Similar conclusions were reached from the calculation of the average and maximum depth of a collection of non-homologous, single-chain globular proteins. These values do not increase indefinitely with the number of residues but rapidly reach a plateau that can be placed between 100 and 200 residues [12] or between 200 and 250 residues [2], depending on the dataset and depth calculation method used. It should be noted that no *a priori* definition of 'domain' is used in these calculations. From the maximum depth value observed, we can also conclude that, despite the scattering of the data, these values are smaller than those that would be expected for perfect spheres or ellipsoids, or, in other words, that the shape of proteins is irregular. For example, a spherical protein of 200 residues would have an expected value for maximum depth of ~20 Å, against a value of ~6 Å and ~12 Å calculated using DPX [4] or the 'nearest water molecule' method [2], respectively, from sets of experimentally determined protein structures. The difference in the calculated values, in turn, arises from the different definition of 'depth' adopted. As discussed previously, the radius of the probe used to identify solvent-accessible atoms in DPX is a further source of variability in calculated depth values.

The residue level
At the residue level, depth values are strictly related to the physicochemical features of the amino acid types and they can therefore be used to monitor and understand a variety of structural properties, from protein thermodynamic stability to the degree of evolutionary conservation. One might expect different amino acid types to have different environmental preferences, and mean residue depths calculated for a dataset of high-resolution structures have actually been used to derive a structure-based classification (Table 1) that is correlated to well known amino acid hydrophobicity scales [12]. The reported mean residue depth values might appear to be rather low, but it should be borne in mind that the number of residues that are totally buried in a protein is relatively small, and this is especially true for small and medium-sized proteins. Furthermore, the method used assigns zero depth to all
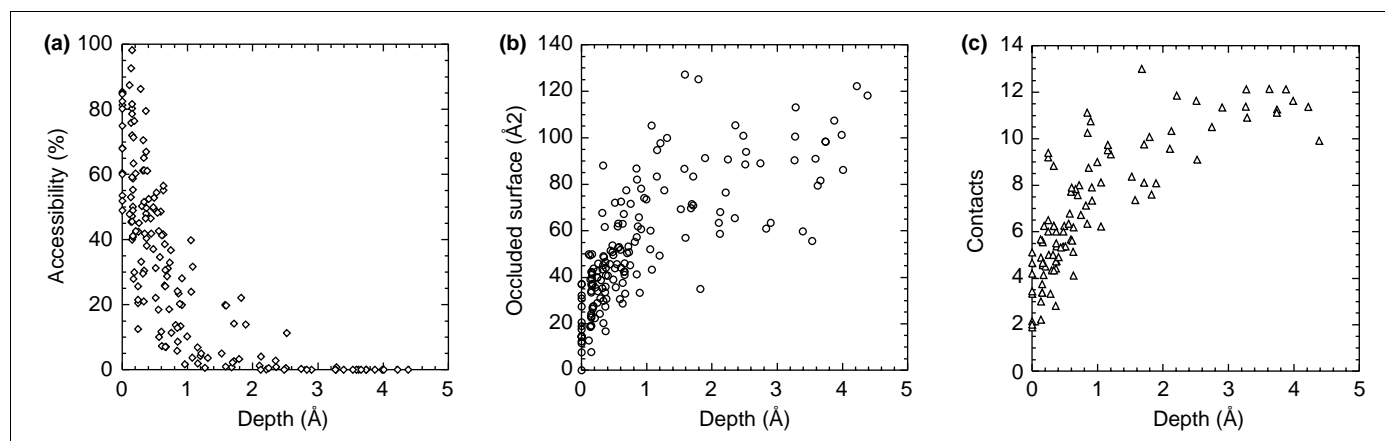

Figure 2. Relationship between depth and other parameters. Residue solvent accessibility **(a)**, residue buried surface **(b)** and mean residue contact number **(c)** for interferon-α2a [7], plotted as a function of mean residue depth. Accessibility (%) was calculated using NACCESS [6], the buried surface (Å²) was calculated using the Occluded Surface (OS) program [11], and the number of atomic contacts was extracted from the OS output and divided by the number of atoms of each amino acid type. Mean residue depth (depth; Å) was calculated using DPX [4].

Table 1. Structure-based mean residue depth for amino acid types[a]

Amino acid	Counts	Depth	sd
A	4086	1.09	0.02
C	940	1.25	0.04
D	3066	0.54	0.01
E	3022	0.48	0.01
F	1994	1.53	0.03
G	3982	0.71	0.02
H	1228	0.85	0.02
I	2784	1.72	0.02
K	3150	0.45	0.01
L	4226	1.52	0.02
M	1012	1.37	0.04
N	2494	0.62	0.01
P	2460	0.68	0.02
Q	1918	0.58	0.01
R	2296	0.61	0.01
S	3410	0.70	0.02
T	3064	0.82	0.02
V	3532	1.57	0.02
W	816	1.38	0.04
Y	1994	1.18	0.02

[a]Amino acid type (single letter code), number of observations (counts) in the dataset of 136 high-resolution structures, mean residue depth (depth, Å) and standard deviation of the mean (sd, Å) [12].

solvent-accessible atoms, and this lowers the average residue values.

The different propensity of an amino acid type to be more or less buried in the protein interior can also be represented as a mutation matrix containing the difference in the mean residue depth for every amino acid couple. Because a larger 'chemical difference', as defined by Grantham [15], has been correlated with a higher probability of observing a disease-related mutation in human genes [16,17], it can be speculated that this matrix could be employed in the prediction of particularly disruptive mutations.

Depth can in fact be correlated with protein stability, as measured experimentally by the free energy of folding, ΔG_{U-F}, for a series of proteins containing cavity-creating mutations. Assuming that the protein structure does not change upon mutation, the difference in side-chain depth between two mutants was calculated, and was shown to correlate better with $\Delta\Delta G_{U-F}$ than with both the number of nearby methylene groups [18] and the difference in the area buried upon folding [2].

Also involved in protein engineering and *de novo* protein design, is the intriguing observation that residue depth and residue conservation are related. From the structural alignment of proteins sharing the same fold type, but displaying no or very limited sequence homology, it was noted that residues that are most deeply buried also display a high degree of conservation [12]. Even though the opposite is not necessarily true, because residue conservation can be linked to functional rather then structural requirements, this correlation opens up interesting perspectives in protein design, evolution and folding. It has been proposed that there is an evolutionary pressure to conserve residues that belong to the 'folding nucleus' [19]. These residues are well conserved because, in addition to stabilization purposes, they play a key role in accelerating the folding process in structurally related

proteins. It remains to be established whether the conservation of 'deep' residues is merely the result of the conservation of the hydrophobic core, or is also related to folding requirements.

Many biological processes involve specific protein–protein interactions, often made by large interacting surface patches. Although the interaction energy is proportional to the hydrophobic surface area buried upon binding, it is often difficult to estimate the contribution of a single residue. From the analysis of protein–protein heterodimeric complexes and their mutants, the change in residue depth upon complex formation has been correlated with the free energy of formation of the complex [2]. Also in this case, the difference in depth correlates better with the stability of the complex than does the accessible area difference, and residue depth can be used to identify residues that contribute significantly to the complex interface. For example, residue depth changes have been employed to identify residues in the human growth hormone binding protein that contribute most to the interaction with the growth hormone [2].

The atomic level
Depth has also turned out to be useful for understanding the structural properties of proteins at the atomic level. It has long been recognized from NMR measurements that deeply buried amide protons exchange slowly with the solvent. The distance between amide N atoms in hen egg-white lysozyme and the nearest hypothetical water molecule was shown to bear a linear correlation with amide H/D exchange rates, as experimentally determined by NMR [1]. Although H/D exchange rates greatly depend also on the formation of hydrogen bonds in secondary structural elements, depth was shown to correlate better than solvent accessibility with amide H/D exchange rates in several proteins, such as thioredoxin, calbindin, cytochrome c-551 and ribonuclease [2].

An area of possible development is atom type classification. Atom typing in proteins has been most often based on the chemical properties of their equivalent atoms in small molecules. A reduction in the number of atom types that can effectively describe the physicochemical properties of a protein could greatly simplify the computational approaches to protein structure prediction and analysis, and to docking simulations [20]. Mean atom depths can thus be calculated and atoms clustered into groups that share the same tendency to be more or less buried in the protein core (O. Carugo and A. Pintar, unpublished).

Another aspect of protein–protein interaction is given by post-translational modifications of proteins, one of the best studied being phosphorylation. With the fast increasing quantity of structural information that is becoming available, structure-, rather then sequence-, based predictions of post-translational modifications are becoming more and more important. Thus, depth has been used to partition atoms into discrete layers in the 3D structures of proteins [12]. Atoms that, although totally buried, are close to the protein surface (within the first buried layer, which includes atoms in the distance range of 0.75–1.75 Å from the surface layer), as measured from atom depth, have been found to be potential targets for phosphorylation.

This finding suggests that these atoms could become solvent accessible through internal dynamics or small conformational transitions, and should thus be taken into account in structure-based predictions of post-translational modification sites.

In conclusion, atom and residue depth is a geometric parameter that can be easily calculated from 3D structures, and can provide valuable information on the protein interior. Unlike other structural descriptors, such as the solvent-accessible surface area, the buried surface area and the number of contacts, depth is able to distinguish atoms that are buried but close to the surface from those that are deep inside into the protein structure, providing a sensitive tool for the analysis of the protein core.

At the molecular level, depth has been used to gain insight into average protein domain size without the need for a preset 'domain' definition. At the residue level, it has been found that depth is correlated with the thermodynamic stability of protein mutants, to the free energy of formation of protein complexes and to amino acid type hydrophobicity. At the atomic level, a good correlation has been found between depth and H/D amide proton exchange rates. The conservation of deep residues at structurally equivalent positions in proteins of the same fold type has generated interesting questions relating to evolution and folding.

From the point of view of future developments, it can be anticipated that both mean atom and mean residue depths can be converted into mean force potentials, to be used in protein structure validation, prediction and folding simulations.

References

1 Pedersen, T.G. *et al.* (1991) A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. *J. Mol. Biol.* 218, 413–426
2 Chakravarty, S. and Varadarajan, R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability. *Structure Fold. Des.* 7, 723–732
3 Connolly, M.L. (1993) The molecular surface package. *J. Mol. Graph.* 11, 139–141
4 Pintar, A. *et al.* (2003) DPX: for the analysis of the protein core. *Bioinformatics* 19, 313–314
5 Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400
6 Hubbard, S.J. and Thornton, J.M. (1993) NACCESS. *Computer program, Department of Biochemistry and Molecular Biology.* University College London
7 Klaus, W. *et al.* (1997) The three-dimensional high resolution structure of human interferon alpha-2a determined by heteronuclear NMR spectroscopy in solution. *J. Mol. Biol.* 274, 661–675
8 Carugo, O. and Bordo, D. (1999) How many water molecules can be detected by protein crystallography? *Acta Crystallogr. D Biol. Crystallogr.* 55, 479–483
9 Raschke, T.M. and Marqusee, S. (1998) Hydrogen exchange studies of protein structure. *Curr. Opin. Biotechnol.* 9, 80–86
10 Ahmad, S. and Gromiha, M.M. (2002) NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 18, 819–824
11 Pattabiraman, N. *et al.* (1995) Occluded molecular surface: analysis of protein packing. *J. Mol. Recognit.* 8, 334–344
12 Pintar, A. *et al.* (2003) Atom depth as a descriptor of the protein interior. *Biophys. J.* 84, 2553–2561
13 Xu, D. and Nussinov, R. (1998) Favorable domain size in proteins. *Fold. Des.* 3, 11–17
14 Wheelan, S.J. *et al.* (2000) Domain size distributions can predict domain boundaries. *Bioinformatics* 16, 613–618
15 Grantham, R. (1974) Amino acid difference formula to help explain protein evolution. *Science* 185, 862–864
16 Krawczak, M. *et al.* (1998) Neighboring-nucleotide effects on the rates of germ-line single-base- pair substitution in human genes. *Am. J. Hum. Genet.* 63, 474–488
17 Cooper, D.N. *et al.* (1998) The human gene mutation database. *Nucleic Acids Res.* 26, 285–287
18 Serrano, L. *et al.* (1992) The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* 224, 783–804
19 Mirny, L. and Shakhnovich, E. (2001) Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* 308, 123–129
20 Tsai, J. *et al.* (2001) Determining the minimum number of types necessary to represent the sizes of protein atoms. *Bioinformatics* 17, 949–956
21 Sayle, R.A. and Milner-White, E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* 20, 374